Vo Declaration
Exhibit M

EXHIBIT C

Message

From: Meta via Workplace [[redacted]@fbworkmail.com]
Sent: 11/21/2023 9:56:02 PM
To: Nikolay Bashlykov [[redacted]@meta.com]
Subject: Shruti Bhosale tagged you in NextGen Working Group.

Shruti Bhosale tagged you in NextGen Working Group.

Shruti Bhosale posted in NextGen Working Group
November 21 at 9:55 PM

NextGen Data Weekly Update (Nov 21, 2023)

Workstream: Have 12T tokens, with the right data mix, and curriculum, that are of high enough quality for us to train Llama3 to GPT4 level.

Links: People & Process

1. HIGHLIGHTS

- We have completed training and applying two quality classifiers to rank RedPyjama-V2 ~17 Trillion GPT-4 tokens. Next, we will start an ablation comparing Top 3T tokens of RedPyjama-V2 with Top 3T tokens of our CC dataset.
- We now experimentally know that Top 3T tokens give better or equivalent results as compared to Top 2T tokens of our new CC dataset. Top 6T tokens (out of 8T tokens) is not high enough quality and leads to some degradation in downstream metrics. We are currently testing using Top 4T tokens.
- We are on track to combine basic heuristic filters, CC training curriculum learnings, best quality classifiers, and best data mix weighting results into the first version of the finalized data mix that can be handed off to the scaling laws project by today or Wednesday morning.
- **Redacted**
- We got approximate parity on English performance for the first time when adding 5% multilingual data to the data mix. Next steps are to combine this with our new 128k vocabulary multilingual tokenizer.

2. LOWLIGHTS

- We have only consistently been able to use ~1024 out of our promised allocation of 1420 H100 gpus. The entitlement allocation of GPUs on EAG is not set correctly. A100s are

2x slower and for experiments where we need to train to 100k-200k+ updates, we have been slowed down.

- Changing the datamix in the middle of training doesn't work well in xlformers and we are working on supporting it. As a result, some of the training curriculum experiments have been blocked or delayed. Punit Singh Koura is now helping to add more flexibility to changing our data mix mid-training.

3. PROGRESS & PRIORITIES

1. **More**

Redacted

2. In collaboration with the Experimental workstream, we are starting to look into getting even more Math data into our datamix (e.g. Proof-Pile-2, scraping from math websites, math magazines etc.).
3. First-party data
    1. Latest stats on raw 1PD data to be cleaned:
        1. FB posts and comments: 260T chars where 76T are English
        2. Speech: 1.9T chars
        3. Business Messaging: 4.7T chars where 0.7 are English
    2. Jacob Xu published first version of FB Post & Comment Analysis. Identified priorities to work on enriching with internal signals, and then start looking into Tier0(Long Post without comments) and Tier1(Long Post with Long comments)
    3. Xuchao Jia came up with FB Post/Comment Dedu Strategy suggesting different treatment with post length groups.
    4. Nikolay Pavlovich Laptev implemented Gopher heuristics in fb prod and computed the statics on 30% of FB posts in only 1.5 hours. The next step is to look closer in the distribution and determine proper filtering thresholds
4. RedPajama2
    1. David Esiobu performed an insightful analysis of quality classifiers score distribution difference comparing new CC v.s. RP2.
        1. The finding is that scores on RP2 docs are a little higher on average. On the same doc RP2 wins v.s. CC-line-dedup about 80% of the time
    2. David Esiobu finished inferencing new LLama classifier for shard=0 of RP2
    3. Punit Singh Koura finished inferencing the wikiref classifier for all shards of RP2 data
    4. Punit Singh Koura finished computing the URL blocklists for all shards of RP2

5. Next step is Punit Singh Koura will start samping the RP2 data and run an experiment to compare with our new CC using the top 1.5T tokens

2. **Better**

1. Multilingual 128k (100k GPT4 + 28k multilingual vocab) Tokenizer Aston Zhang Chloe Bi

1. We are switching from the GPT-4 tokenizer to the new cl_toplang_128k tokenizer for all our data and scaling-law pre-training runs of Llama 3. Aston Zhang has shown that,

1. Our new tokenizer has a comparable training loss with the GPT-4 tokenizer
2. Our new tokenizer has comparable eval results with the GPT-4 tokenizer on English benchmarks
3. Our new tokenizer has better evaluation results than the GPT-4 tokenizer on multilingual benchmarks
4. Our new tokenizer has better compression ratios (#tokens/#chars) than the GPT-4 tokenizer without affecting that of code or English
5. More details could be found in this note

2. Heuristics

1. First ablation experiment of rule-based heuristics filtering is completed by Frank Zhang. This initial version filters out bad documents based on 10 curated rules such as removing docs containing > 300 dirty words, removing docs whose mean word length > 20, etc.

1. Experiment results show decent improvement on Knowledge and Reasoning, slightly regressed on code and math
2. The next step is to test not filtering out docs contain curly brackets and see whether we can improve on code and math

2. Viktor Kerkez trained a new classifier based on human annotated dataset using selected heuristics features

1. The accuracy of this new version improves compared to the previous one thanks to more human labeled data and heuristics calculated on latest line deduped text corpus
2. The next step is to finish computing the heuristics on new CC shard0 and start an ablation experiment to test the effectiveness of the classifier

3. Token distribution KL-divergence

1. Andrew Poulton was able to implement the calculation on Spark and finished over Edouard_cc in ~2.5 hrs. The next step is to scale to 100% on new CC
2. Nikolay Bashlykov finished processing libgen and removed ~0.4-1% of out-of-distribution documents. Filtered examples include copyright lines, emails from

text and some OCR repetition artifacts. Next step is to run an ablation experiment to test the effect of the filtering

3. Continuing to refine quality classifier for ranking common crawl:

1. Completed running prompt v2 distillberta classifier on all cleaned, de-duped docs (8T tokens) on the AWS cluster

4. We now experimentally know that Top 3T tokens give better or equivalent results as compared to Top 2T tokens of our new CC dataset. Top 6T tokens (out of 8T tokens) is not high enough quality and leads to some degradation in downstream metrics. We are currently testing using Top 4T tokens.

1. Some training curriculum variants are blocked as xlformers doesn't yet support changing the data mix mid-training and Punit Singh Koura is working to add support for that.

2. Todor Mihaylov is helping prepare Top 3T, 4T variants with the best two quality classifiers we have so far. And we will start more ablations with these.

5. NSFW Detection

1. We are detecting NSFW content in Libgen-fiction with the aim of filtering it out or heavily downsampling it.

6. Better PII filtering

1. Yuchen Zhang Eric Smith are helping us get better PII filtering rules for our English data.

7. Multilingual

1. For the first time, we see approximate parity and no degradation for English when adding 5% multilingual for 100k steps with the best llama3 data mix so far. Chloe Bi

2. Experiment with the new multilingual tokenizer (100k GPT4 + 28k multilingual merged) Chloe Bi

3. More multilingual CC preprocessing and cleaning Chloe Bi Jacob Xu

3. **Faster**

1. Todor Mihaylov did an amazing knowledge sharing of the Llama3 CC ranking and sampling pipeline. Video recording could be found here.

2. Working with Symon Perriman to see if we can use inference GPUs for data processing.

3. With DE/DI support, we are continuing WIP to migrate data processing pipelines to production. This is unlocking more contributors to infra and processing work (GenAI media foundations team for 1PD work, AI infra data infra teams for 1PD/common crawl).

4. TEAM

- Several of the team will be taking longer leaves/PTO in December, so we are planning their priorities carefully, expanding documentation, and making sure there is knowledge redundancy of any tasks those teammates owned.

Like Comment Share

View on Workplace